CONFERENCIA: Visualising patterns in text

Michael O'Donnell (Universidad Autónoma de Madrid)

RESUMEN:

En la mayoría de estudios de corpus se emplean estadísticas de algún tipo para llegar a conclusiones, por ejemplo, mediante la comparación de la ocurrencia relativa de palabras, sintagmas, o estructuras sintácticas en diferentes corpus. Sin embargo, los números por sí mismos no son comprendidos por los humanos con facilidad, sino que entendemos más fácilmente los datos que se presentan en forma visual, tales como tablas, gráficos, etc.

Más que explorar pautas en los corpus, esta conferencia se centra en explorar pautas en textos aislados de una forma visual. Por ejemplo, las Nubes de Palabras se han utilizado en los últimos años para presentar los elementos léxicos más sobresalientes o *salientes* en el texto, mostrando únicamente las palabras más sobresalientes, y aumentando el tamaño de la palabra cuanto más prominente es.

En la primera parte de la conferencia se presenta un modo alternativo de visualización en el que, en lugar de mostrar las palabras fuera de su contexto, se muestra el propio texto con las palabras en diferente color y tamaño en función de su prominencia o *saliencia*.

La segunda parte de la conferencia se centra en la visualización de pautas sintácticas en el texto. Los recientes avances en la tecnología de análisis sintáctico han hecho que actualmente dispongamos de analizadores sintácticos razonablemente fiables para muchas lenguas. No obstante, la visualización de un texto en forma de árboles sintácticos no nos ayuda a percibir qué pautas hay en el texto. Las estadísticas en crudo, tales como el número de cláusulas modales, pasivas, etc., que hay en el texto tampoco ayudan. Presentaremos diferentes maneras de visualizar el texto de tal modo que la naturaleza cambiante de las pautas sintácticas se pueda hacer visible para el analista.

Palabras clave: Visualización de datos textuales, Nubes de Palabras, pautas léxicas, pautas sintácticas

SUMMARY:

In most corpus studies, statistics of some kind are used to reach some conclusions, for instance, comparing the relative occurrence of words, phrases, or syntactic structures in different corpora. However, numbers themselves are not so readily understood by humans, we more readily comprehend visually-presented data, such as bar charts, pie charts, etc.

Rather than exploring patterns in corpora, this talk focuses on exploring patterns within individual texts in a visual way. For instance, Word Clouds have been used in recent years to present the most salient lexical items in a text, with the most salient words only shown, and with the size of a word increasing with its salience.

In the first part of this talk, an alternative way of presenting lexical salience is presented: rather than showing words out of their context, the text itself is presented with words shaded or sized according to salience.

The second part of the talk focuses on visualising syntactic patterns in text. Recent advances in parsing technology mean that we have available reasonably reliable syntactic parsing for many languages. However, viewing a text in terms of a series of syntactic trees does not help us perceive the patterns in the text. Raw statistics, such as the number of modal clauses, passives, etc. in the text also does not help. We will present various ways to visualise a text such that the changing nature of syntactic patterns throughout the text become visible to the analyst.

Key words: Textual data visualising, Word Clouds, lexical patterns, syntactic patterns